

# Nonparametric ROC Summary Statistics for Correlated Diagnostic Marker Data

Liansheng Larry Tang<sup>1</sup>, Aiyi Liu<sup>2</sup>, Zhen Chen<sup>2</sup>, Enrique F. Schisterman<sup>2</sup>, Bo Zhang<sup>3</sup>, and Zhuang Miao<sup>1</sup>

We propose efficient nonparametric statistics to compare medical imaging modalities in multi-reader multi-test data and to compare markers in longitudinal ROC data. The proposed methods are based on the weighted area under the ROC curve which includes the area under the curve and the partial area under the curve as special cases. The methods maximize the local power for detecting the difference between imaging modalities. The asymptotic results of the proposed methods are developed under a complex correlation structure. Our simulation studies show that the proposed statistics result in much better powers than existing statistics. We applied the proposed statistics to an endometriosis diagnosis study. Copyright © 2010 John Wiley & Sons, Ltd.

**Keywords:** ROC curve; Optimal weights; Wilcoxon statistics; Correlated data

## 1. Introduction

In medical imaging studies, one is concerned about whether a newly developed imaging modality is more accurate than traditional modalities to correctly discriminate a subject with abnormal lesions from a subject without such lesions. Imaging modalities are considered as an example of diagnostic markers, which are used to distinguish a subject with a particular condition (“the diseased”) from a subject without the condition (“the non-diseased”). For diagnostic markers that generate binary test results, their accuracy can be summarized in terms of sensitivity (probability of identifying a diseased subject when the disease truly exists) and specificity (probability of correctly ruling out a non-diseased subject when the disease is truly absent). For diagnostic markers that generate discrete or continuous test results, the receiver operating characteristic (ROC) curve is a standard statistical tool to describe and compare the accuracy of markers [1]. The ROC curve combines all possible pairs of sensitivities and 1 – specificities from different decision thresholds and thus describes the accuracy of markers apart from decision thresholds.

For correlated results from two diagnostic markers, parametric and nonparametric methods have been proposed to compare ROC summary measures. Parametric methods for the area under the curve (AUC) assume distributions (e.g. negative exponential, normal, lognormal, gamma) on marker measurements [2, 3]. These methods may not perform

<sup>1</sup> Department of Statistics, George Mason University, Fairfax, VA 22030, USA

<sup>2</sup> Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, Rockville, Maryland 20852, USA

<sup>3</sup> School of Biological and Population Health Sciences, College of Public Health and Human Sciences, Oregon State University, Corvallis, 97331, USA

\* Correspondence to: ltang1@gmu.edu, Department of Statistics, George Mason University, Fairfax, VA 22030, USA

# Statistics in Medicine

well if the parametric assumptions are invalid. The semiparametric ROC estimation based on the logistic regression is proposed by [4]. As an alternative, nonparametric methods do not require distribution assumptions and are robust to model misidentification. Nonparametric methods to estimate and compare two AUCs have been proposed by [5], [6], and others. These methods are based on results for U-statistics because an empirical AUC statistic is essentially a Wilcoxon rank sum statistic [7]. However, if two ROC curves intersect, their AUCs may be equal and do not provide valid information for the comparison. Moreover, summarizing the entire ROC curve may include irrelevant information about the marker's accuracy when one is only interested in some range of specificities. For example, acceptable specificities are high for early cancer detection tests. The partial area under the curve (pAUC), which summarizes part of the ROC curve in the range of desired specificities, may be a better alternative. Nonparametric methods to compare pAUCs are proposed by [8]. **Utilizing the pAUCs is particularly important in comparing markers which are developed to screen a large population for certain diseases, for example, breast cancer [9]. A lower specificity for a large population leads to many more falsely classified non-diseased subjects who may have to undergo a more invasive test subsequently. It is thus desired to compare screening markers at a higher range of specificities.**

In this paper we propose efficient nonparametric ROC statistics to analyze multi-reader multi-test ROC data and to nonparametrically summarize correlated longitudinal ROC data. The proposed method not only includes many nonparametric ROC summary measures as special cases, but also maximizes the local power for detecting the difference between markers. The rest of the article is organized as follows. In Section 2 we introduce the new statistics for multi-reader multi-test ROC data and longitudinal ROC data, and discuss the equivalence between our statistics and the generalized Wilcoxon statistics under specific assumptions. Section 3 gives the variance expressions for the proposed statistics. Section 4 reports simulation results to illustrate the small sample performance of the proposed ROC statistics and their theoretical variances. Section 5 applies the proposed method to a real example on the diagnosis of endometriosis. Section 6 gives some discussion.

## 2. Methods

### 2.1. Definition of nonparametric ROC summary statistics

We first define some notations. Suppose test result  $X_{\ell ip}$  of marker  $\ell$  is from the  $p$ th abnormal location in the diseased subject  $i$ , where  $\ell = 1, \dots, L$ ,  $p = 0, 1, \dots, m_{\ell i}$ , and  $i = 1, \dots, M$ . Test result  $Y_{\ell jq}$  of marker  $\ell$  is from the  $q$ th normal location in the non-diseased subject  $j$ , where  $\ell = 1, \dots, L$ ,  $q = 0, 1, \dots, n_{\ell j}$ , and  $j = 1, \dots, J$ . Here the total number of subjects is  $N = M + J$ . The joint pairwise cumulative function of  $(X_{\ell_1 ip_1}, X_{\ell_2 ip_2})$  is taken to be  $S_{D, \ell_1, \ell_2}(x_1, x_2)$ ,  $p_1, p_2 = 1, \dots, m_{\ell i}$ , with marginal survival functions  $X_{\ell ip} \sim S_{D, \ell}(x)$ . Similarly we define  $(Y_{\ell_1 jq_1}, Y_{\ell_2 jq_2}) \sim S_{\bar{D}, \ell_1, \ell_2}(y_1, y_2)$ ,  $q_1, q_2 = 1, \dots, n_{\ell j}$ , with marginal survival functions  $Y_{\ell jq} \sim S_{\bar{D}, \ell}(y)$ . The ROC curve for the  $\ell$ th marker is then given by  $ROC_{\ell}(u) = S_{D, \ell} \{S_{\bar{D}, \ell}^{-1}(u)\}$ , where the false positive rate (FPR)  $u$  is in  $[0, 1]$ . The resulting  $\ell$ th weighted area under the curve (wAUC) is

$$\Omega_{\ell} = \int_0^1 S_{D, \ell} \{S_{\bar{D}, \ell}^{-1}(u)\} dW(u), \quad (1)$$

with a probability measure  $W(u)$  defined on  $u$ , for  $u \in [0, 1]$ . Included in this class of accuracy measures are AUC, pAUC between FPRs  $u_1$  and  $u_2$ , and the sensitivity at a given level of FPR  $u_0$ .  $W(u)$  can also be defined as certain distribution functions, such as the beta cdf, to assign varying weight to the specificity. The detailed discussion is in [10].

By substituting the functions  $S_{D, \ell}$  and  $S_{\bar{D}, \ell}$  with their respective empirical function  $\hat{S}_{D, \ell}$  and  $\hat{S}_{\bar{D}, \ell}$ , the nonparametric wAUC estimator is given by  $\hat{\Omega}_{\ell} = \int_0^1 \hat{S}_{D, \ell} \{\hat{S}_{\bar{D}, \ell}^{-1}(u)\} dW(u)$ . The empirical survival functions  $\hat{S}_{D, \ell}$  and  $\hat{S}_{\bar{D}, \ell}$  are defined

by

$$\begin{aligned}\hat{S}_{D,\ell}(x) &= \frac{1}{\sum_{i=1}^M m_{\ell i}} \sum_{i=1}^M \sum_{p=1}^{m_{\ell i}} I(X_{\ell ip} > x), \\ \hat{S}_{\bar{D},\ell}(x) &= \frac{1}{\sum_{j=1}^J n_{\ell j}} \sum_{j=1}^J \sum_{q=1}^{n_{\ell j}} I(Y_{\ell jq} > x).\end{aligned}\quad (2)$$

Denote  $\Omega = (\Omega_1, \Omega_2, \dots, \Omega_L)$ . By substituting  $\hat{S}_{D,\ell}$  and  $\hat{S}_{\bar{D},\ell}$  in Equation (1), the nonparametric estimator of  $\Omega$  is given by  $\hat{\Omega} = (\hat{\Omega}_1, \hat{\Omega}_2, \dots, \hat{\Omega}_L)$ .

We define  $W(u) = u$  for  $0 < u < 1$  to obtain the nonparametric AUC estimator for the  $\ell$ th marker as follows

$$\hat{\Omega}_{A,\ell} = \frac{1}{\sum_{i=1}^M m_{\ell i} \sum_{j=1}^J n_{\ell j}} \sum_{i=1}^M \sum_{p=1}^{m_{\ell i}} \sum_{j=1}^J \sum_{q=1}^{n_{\ell j}} I(X_{\ell ip} > Y_{\ell jq}). \quad (3)$$

The AUC statistic in (3) takes the form of the Wilcoxon rank-sum statistic. It essentially compares the measurements of abnormal locations with those of normal locations. To calculate this statistic, we obtain every possible pair of measurements from an abnormal location and a normal location. We assign 1 if the abnormal location's measurement is larger than the normal location in the pair, and 0 otherwise.  $\hat{\Omega}_{A,\ell}$  is then calculated by averaging the 1's and 0's over all possible pairs. Since the location within each subject is viewed as the unit of sampling, the inference based on the regular Wilcoxon rank-sum statistic is not valid here.

When  $W(u) = (u - u_1)/(u_2 - u_1)$  for  $0 < u_1 \leq u \leq u_2 < 1$ ,  $\hat{\Omega}_\ell$  empirically estimates the partial AUC (pAUC), and its explicit form is given by

$$\frac{1}{\sum_{i=1}^M m_{\ell i} \sum_{j=1}^J n_{\ell j}} \sum_{i=1}^M \sum_{p=1}^{m_{\ell i}} \sum_{j=1}^J \sum_{q=1}^{n_{\ell j}} I(X_{\ell ip} > Y_{\ell jq} | Y_{\ell jq} \in (\hat{S}_{\bar{D},\ell}^{-1}(u_2), \hat{S}_{\bar{D},\ell}^{-1}(u_1))). \quad (4)$$

The pAUC statistic in (4) uses all measurements from the abnormal locations. Since the pAUC is specified to be in the range of  $(u_1, u_2)$ , only measurements from the normal locations which fall in  $(\hat{S}_{\bar{D},\ell}^{-1}(u_2), \hat{S}_{\bar{D},\ell}^{-1}(u_1))$  are used in (4). That is, we sort all measurements from the normal locations from the smallest to the largest, and obtain the order statistics  $Y_{[(1-u_2) \sum_{j=1}^J n_{\ell j}]}$  and  $Y_{[(1-u_1) \sum_{j=1}^J n_{\ell j}]}$ , where  $[x]$  denotes the smallest integer greater than or equal to  $x$ . We then calculate the Wilcoxon rank-sum like statistic by comparing all X's with Y's which are between  $Y_{[(1-u_2) \sum_{j=1}^J n_{\ell j}]}$  and  $Y_{[(1-u_1) \sum_{j=1}^J n_{\ell j}]}$ . The pAUC statistic is useful in disease screening when a high FPR would lead to a large number of falsely diagnosed subjects. It is desirable to evaluate and compare the marker accuracy at the low FPRs rather than the entire range of FPRs. When we are interested in the sensitivity of the  $\ell$ th marker at a particular threshold, say  $c$ , we can specify the probability measure to be a point mass at  $u_0 = S_{\bar{D},\ell}(c)$ . The estimator  $\hat{\Omega}_\ell$  then becomes

$$\frac{1}{\sum_{i=1}^M m_{\ell i}} \sum_{i=1}^M \sum_{p=1}^{m_{\ell i}} I(X_{\ell ip} > Y_{[(1-u_0) \sum_{j=1}^J n_{\ell j}]}). \quad (5)$$

The estimator in (5) is obtained by comparing all X's with  $Y_{[(1-u_0) \sum_{j=1}^J n_{\ell j}]}$ .

In the following sections, we propose efficient nonparametric methods based on the nonparametric estimator of  $\Omega$  to evaluate and compare multiple markers in multi-reader multi-test ROC Data and longitudinal ROC data.

## 2.2. Multi-reader multi-test ROC data

**One type of complex marker data arise frequently in medical imaging studies when radiological images of a patient are evaluated by several radiologists. [11] consider a mixed-effect ANOVA model while allowing for correlation**

among AUC estimators. Their model requires a specific covariance structure among the AUCs. [12] propose a pseudo-generalized estimating equation method and derive large sample theory for the estimators. Their method remains valid under the working-independence assumption.

In a multi-reader multi-test ROC study, suppose the radiologist  $r$ ,  $r = 1, \dots, R$ , rates images for  $M$  diseased subjects and  $J$  non-diseased subjects from  $L$  imaging devices. A radiologist can give one or more ratings to suspicious locations in each subject, that is,  $m_{\ell i}, n_{\ell j} \geq 1$ . We consider  $L = 2$ . Denote  $\Omega_1, \dots, \Omega_R$  as wAUCs from  $R$  readers for modality 1,  $\Omega_{R+1}, \dots, \Omega_{2R}$  as wAUCs from  $R$  readers for modality 2. Common nonparametric approaches for comparing imaging modalities take the difference  $\Omega_r - \Omega_{R+r}$  between two devices for reader  $r$ , and then average these differences over all reader [13]. We can see that such methods are a special case of the linear combination of the weighted AUC statistics for reader-modality combinations. Rather than the simple average of all  $\Omega_r - \Omega_{R+r}$ 's, we propose to use the following weighted linear combination to possibly achieve a higher power to compare markers

$$\Delta_m = \left( \sum_{r=1}^R w_r \right)^{-1} \sum_{r=1}^R [w_r (\Omega_r - \Omega_{R+r})], \quad (6)$$

with positive and bounded weights  $\tilde{W} = (w_1, w_2, \dots, w_R)'$ . The parameter  $\Delta_m$  can be empirically estimated by

$$\hat{\Delta}_m = \left( \sum_{r=1}^R w_r \right)^{-1} \sum_{r=1}^R [w_r (\hat{\Omega}_r - \hat{\Omega}_{R+r})],$$

which compares two modalities with multiple readers.

Various choices of weights exist in the ROC literature.  $\tilde{W}$  may not depend on the data. For instance, if all readers are assumed to be homogeneous with regard to their accuracy of rating images, an equal weight  $w_r = 1/R$  can be assigned to reader  $r$ ,  $r = 1, \dots, R$ . Then with  $m_{\ell i} = n_{\ell j} = 1$  and  $W(u) = 1$  at  $0 < u < 1$ ,  $\hat{\Delta}_m$  becomes the AUC statistic in [13]. When one has to estimate  $\tilde{W}$  from the data, the consistency of estimated weights  $\hat{W}$  in probability is required for the derivation. For instance, a set of optimal weights is introduced by [14] and further developed by [15], who argues that when readers' experience vary greatly, using equal weights may yield a biased AUC estimate. Let the  $R \times R$  covariance matrix of estimated AUC differences,  $(\hat{\Omega}_1 - \hat{\Omega}_{R+1}, \dots, \hat{\Omega}_R - \hat{\Omega}_{2R})'$ , be  $\Sigma_A$ , and its consistent estimator  $\hat{\Sigma}_A$ . They then choose  $\tilde{W} = \hat{\Sigma}_A^{-1} \mathbf{1}$  to obtain a consistent estimator for the AUC difference, where  $\mathbf{1}$  is a  $R$ -dimensional vector of one's. [14] and [15] show that this set of weights are optimal since they maximize the local power to detect the AUC difference between imaging modalities. It is clear that by combining these weights with  $m_{\ell i} = n_{\ell j} = 1$  and  $W(u) = 1$  at  $0 < u < 1$ ,  $\hat{\Delta}_m$  becomes [15]'s statistic. To properly calculate the weights for the proposed statistic, we need to obtain the covariance matrix  $\Sigma$  of  $\hat{\Omega} = (\hat{\Omega}_1, \dots, \hat{\Omega}_{2R})'$ . Since in practice  $\Sigma$  is unknown, its consistent estimator  $\hat{\Sigma}$  can be obtained using the explicit expression (A.1) derived in the Appendix. Since  $\Sigma$  and  $\Sigma_A$  is related via

$$\Sigma_A = \Sigma \mathbf{A},$$

where the  $r$ th column of the  $2R \times R$  matrix  $\mathbf{A}$  has 1's at  $r$ th and  $(R+r)$ th rows and 0 at other rows, the estimated weights are given by

$$\hat{W} = \hat{\Sigma}^{-1} \mathbf{A} \mathbf{1}. \quad (7)$$

### 2.3. Longitudinal biomarker data

Another example of complex marker data comes from longitudinal studies when marker measurements are taken at several times during the studies. Most methodology for longitudinal ROC data rely on appropriate assumptions on the distributions of marker measurements [16]. In longitudinal ROC data, suppose  $L$  markers are measured on  $M$  diseased patients and  $J$  non-diseased patients at times  $t_1, t_2, \dots, t_K$ .

Suppose each subject is repeatedly measured for every marker at each time. Let  $X_{\ell ipk}$  denote the test result of marker  $\ell$  in the  $p$ th repetition on the diseased subject  $i$  at time  $t_k$ , where  $\ell = 1, \dots, L$ ,  $p = 1, \dots, m_{\ell ik}$ ,  $i = 1, \dots, M$ , and  $k = 1, \dots, K$ . Let  $Y_{\ell jqk}$  denote test result of  $\ell$ th marker on the  $p$ th repetition in the non-diseased subject  $j$  at time  $t_k$ , where  $\ell = 1, \dots, L$ ,  $q = 1, \dots, n_{\ell jk}$ ,  $j = 1, \dots, J$ , and  $k = 1, \dots, K$ . The nonparametric wAUC estimator for the  $\ell$ th marker is then given by  $\hat{\Omega}_\ell = \int_0^1 \hat{S}_{D,\ell} \{ \hat{S}_{\bar{D},\ell}^{-1}(u) \} dW(u)$ , where  $\hat{S}_{D,\ell}$  and  $\hat{S}_{\bar{D},\ell}$  are defined by

$$\hat{S}_{D,\ell}(x) = \frac{1}{\sum_{i=1}^M \sum_{k=1}^K m_{\ell ik}} \sum_{i=1}^M \sum_{k=1}^K \sum_{p=1}^{m_{\ell ik}} I(X_{\ell ipk} > x),$$

and  $\hat{S}_{\bar{D},\ell}(x) = \frac{1}{\sum_{j=1}^J \sum_{k=1}^K n_{\ell jk}} \sum_{j=1}^J \sum_{k=1}^K \sum_{q=1}^{n_{\ell jk}} I(Y_{\ell jqk} > x).$  (8)

By defining  $W(u)$  accordingly in the wAUC estimator, we obtain the nonparametric AUC estimator for the  $\ell$ th marker:

$$\frac{1}{\sum_{i=1}^M \sum_{k=1}^K m_{\ell ik} \sum_{j=1}^J \sum_{k=1}^K n_{\ell jk}} \sum_{i=1}^M \sum_{k=1}^K \sum_{p=1}^{m_{\ell ik}} \sum_{j=1}^J \sum_{k_2=1}^K \sum_{q=1}^{n_{\ell jk_2}} I(X_{\ell ipk_1} > Y_{\ell jqk_2}),$$

the partial AUC estimator:

$$\frac{\sum_{i=1}^M \sum_{k_1=1}^K \sum_{p=1}^{m_{\ell ik_1}} \sum_{j=1}^J \sum_{k_2=1}^K \sum_{q=1}^{n_{\ell jk_2}} I(X_{\ell ipk_1} > Y_{\ell jqk_2} | Y_{\ell jqk_2} \in (\hat{S}_{\bar{D},\ell}^{-1}(u_2), \hat{S}_{\bar{D},\ell}^{-1}(u_1)))}{\sum_{i=1}^M \sum_{k=1}^K m_{\ell ik} \sum_{j=1}^J \sum_{k=1}^K n_{\ell jk}},$$

and the sensitivity estimator at the FPR of  $u_0$ ,

$$\frac{1}{\sum_{i=1}^M \sum_{k=1}^K m_{\ell ik}} \sum_{i=1}^M \sum_{k=1}^K \sum_{p=1}^{m_{\ell ik}} I(X_{\ell ipk} > Y_{[(1-u_0) \sum_{j=1}^J \sum_{k=1}^K n_{\ell jk}]}).$$

We define  $h$  to be a real-valued function of  $\hat{\Omega}$ . Here the function  $h$  is defined on  $\mathbb{R}^L$ , and has continuous partial derivatives of order 2. Let the ROC summary measure be  $\Delta_h = h(\hat{\Omega})$ . Its empirical estimator is given by

$$\hat{\Delta}_h \equiv h(\hat{\Omega}) = h \left( \int_0^1 \hat{S}_{D,1} \{ \hat{S}_{\bar{D},1}^{-1}(u) \} dW(u), \dots, \int_0^1 \hat{S}_{D,L} \{ \hat{S}_{\bar{D},L}^{-1}(u) \} dW(u) \right). \quad (9)$$

The statistic above can be used to compare two longitudinal markers when  $h$  is a linear contrast.  $\hat{\Delta}_h$  also includes a broad range of ROC statistics. It is the weighted AUC statistic in [17] and later in [10] for evaluating and comparing markers. When  $W(u) = 1$  at  $0 < u < 1$  and  $h$  is a linear function,  $\hat{\Delta}_h$  is the generalized AUC statistic in [13]. When  $W(u) = 1$  at  $0 < u < 1$ ,  $\hat{\Delta}_h$  is the AUC statistic in [18], assuming no correlation between  $X$  and  $Y$ , which allows for multiple observations per patient from each marker. When  $W(u) = (u - a)/(b - a)$  for  $0 < a < u < b < 1$  and  $h(\Omega_1, \Omega_2) = \Omega_1 - \Omega_2$ ,  $\hat{\Delta}_h$  is the pAUC statistic in [8] for comparing two markers.

**When there are two longitudinal markers in the study, the optimal combination for comparing the two markers can be obtained using the similar steps in the aforementioned multi-reader multi-test studies. Suppose  $L = 2$ . Let  $\Omega_{\ell,k}$  be the wAUC of marker  $\ell$ ,  $\ell = 1, 2$ , at time  $t_k$  and  $\hat{\Omega}_{\ell,k}$  be its nonparametric estimator given by  $\hat{\Omega}_{\ell,k} = \int_0^1 \hat{S}_{D,\ell,k} \{ \hat{S}_{\bar{D},\ell,k}^{-1}(u) \} dW(u)$ , where  $\hat{S}_{D,\ell,k}$  and  $\hat{S}_{\bar{D},\ell,k}$  are defined by**

$$\hat{S}_{D,\ell,k}(x) = \frac{1}{\sum_{i=1}^M m_{\ell ik}} \sum_{i=1}^M \sum_{p=1}^{m_{\ell ik}} I(X_{\ell ipk} > x), \text{ and } \hat{S}_{\bar{D},\ell,k}(x) = \frac{1}{\sum_{j=1}^J n_{\ell jk}} \sum_{j=1}^J \sum_{q=1}^{n_{\ell jk}} I(Y_{\ell jqk} > x). \quad (10)$$

**Note that the estimation of  $\Omega_{\ell,k}$  is based on every individual time point. One can take difference of the wAUCs of**

two markers, and simply average these differences over all time points. We may also use the following weighted linear combination to possibly achieve a higher power to compare markers

$$\Delta_\ell = \left( \sum_{k=1}^K w_k \right)^{-1} \sum_{k=1}^K [w_k (\Omega_{1,k} - \Omega_{2,k})], \quad (11)$$

with positive and bounded weights  $\tilde{W} = (w_1, w_2, \dots, w_K)'$ . The parameter  $\Delta_\ell$  can be empirically estimated by

$$\hat{\Delta}_\ell = \left( \sum_{k=1}^K w_k \right)^{-1} \sum_{k=1}^K [w_k (\hat{\Omega}_{1,k} - \hat{\Omega}_{2,k})].$$

Similarly as in the previous section, the  $2K \times 2K$  covariance matrix  $\Sigma$  of  $\hat{\Omega} = (\hat{\Omega}_{1,k}, \dots, \hat{\Omega}_{2K})'$  can be estimated can be obtained using the explicit expression in (A.1). Thus the estimated weights are given by the same expression as (7).

### 3. Asymptotic variance expressions of the proposed statistics

In this section we derive the asymptotic variances for the proposed statistics in the multi-reader multi-test data and the longitudinal data. We first show the explicit variance expressions for  $\hat{\Delta}_m$ , and then show the variance expression for the more general statistic  $\hat{\Delta}_h$  in (9) for the longitudinal data.

The numbers of abnormal locations within a diseased subject may differ, and so are the numbers of normal locations within a non-diseased subject. Denote  $\tilde{m}_\ell = \sum_{i=1}^M m_{\ell i}$ , and  $\tilde{n}_\ell = \sum_{j=1}^J n_{\ell j}$ . Assume that  $S_{D,\ell}$  and  $S_{\bar{D},\ell}$  have continuous and positive derivatives,  $S'_{D,\ell}$ , and  $S'_{\bar{D},\ell}$ . In Appendix we show that the proposed statistic,  $\hat{\Delta}_m$ , for the multi-reader multi-test ROC data is asymptotically normal when sample sizes are large. The variance of  $\hat{\Delta}_m$  has the following expression when sample sizes get large:

$$\text{var}(\hat{\Delta}_m) = \tilde{v}_X + \tilde{v}_Y, \quad (12)$$

with

$$\begin{aligned} \tilde{v}_X = & \frac{1}{M \tilde{m}_{\ell_1} \tilde{m}_{\ell_2} (\sum_{r=1}^R w_r)^2} \sum_{1 \leq \ell_1, \ell_2 \leq 2R} \sum_{i=1}^M \tilde{m}_{\ell_1 i} \tilde{m}_{\ell_2 i} (-1)^{I(\ell_1, \ell_2)+1} \left( \iint [S_{D, \ell_1, \ell_2} \{S_{\bar{D}, \ell_1}^{-1}(s), S_{\bar{D}, \ell_2}^{-1}(t)\} \right. \\ & \left. - S_{D, \ell_1} \{S_{\bar{D}, \ell_1}^{-1}(s)\} S_{D, \ell_2} \{S_{\bar{D}, \ell_2}^{-1}(t)\}] dW(s) dW(t) \right), \end{aligned}$$

and

$$\begin{aligned} \tilde{v}_Y = & \frac{1}{M \tilde{n}_{\ell_1} \tilde{n}_{\ell_2} (\sum_{r=1}^R w_r)^2} \sum_{1 \leq \ell_1, \ell_2 \leq 2R} \sum_{j=1}^J \tilde{n}_{\ell_1 j} \tilde{n}_{\ell_2 j} (-1)^{I(\ell_1, \ell_2)+1} \left( \iint r_{\ell_1}(s) r_{\ell_2}(t) \right. \\ & \left. \times [S_{\bar{D}, \ell_1, \ell_2} \{S_{D, \ell_1}^{-1}(s), S_{D, \ell_2}^{-1}(t)\} - st] dW(s) dW(t) \right), \end{aligned}$$

where  $I(\ell_1, \ell_2) = 1$ , if  $|\ell_2 - \ell_1| < R$ , and 0, otherwise, and

$$r_\ell(u) = S'_{D,\ell} \{S_{\bar{D},\ell}^{-1}(u)\} / S'_{\bar{D},\ell} \{S_{D,\ell}^{-1}(u)\}, \quad \text{for } \ell = 1, \dots, L.$$

The marginal and joint survivor functions can also be empirically estimated.

Denote  $m_\ell = \sum_{i=1}^M \sum_{k=1}^K m_{\ell i k}$ , and  $n_\ell = \sum_{j=1}^J \sum_{k=1}^K n_{\ell j k}$ . we show in Appendix that the proposed statistic,  $\hat{\Delta}_h$  in (9) for the longitudinal data is also asymptotically normal, and the variance of  $\hat{\Delta}_h$  takes on the following form when



sample sizes are large,

$$\text{var}(\hat{\Delta}_h) = v_X + v_Y, \quad (13)$$

where

$$v_X = \frac{1}{M m_{\ell_1} m_{\ell_2}} \sum_{i=1}^M m_{\ell_1 i} m_{\ell_2 i} \frac{\partial h}{\partial \Omega_{\ell_1}} \frac{\partial h}{\partial \Omega_{\ell_2}} \left( \iint [S_{D, \ell_1, \ell_2} \{S_{\bar{D}, \ell_1}^{-1}(s), S_{\bar{D}, \ell_2}^{-1}(t)\} - S_{D, \ell_1} \{S_{\bar{D}, \ell_1}^{-1}(s)\} S_{D, \ell_2} \{S_{\bar{D}, \ell_2}^{-1}(t)\}] dW(s) dW(t) \right),$$

and

$$v_Y = \frac{1}{M n_{\ell_1} n_{\ell_2}} \sum_{j=1}^J n_{\ell_1 j} n_{\ell_2 j} \frac{\partial h}{\partial \Omega_{\ell_1}} \frac{\partial h}{\partial \Omega_{\ell_2}} \left( \iint r_{\ell_1}(s) r_{\ell_2}(t) [S_{\bar{D}, \ell_1, \ell_2} \{S_{\bar{D}, \ell_1}^{-1}(s), S_{\bar{D}, \ell_2}^{-1}(t)\} - st] dW(s) dW(t) \right),$$

where

$$r_{\ell}(u) = S'_{D, \ell} \{S_{\bar{D}, \ell}^{-1}(u)\} / S'_{\bar{D}, \ell} \{S_{\bar{D}, \ell}^{-1}(u)\}, \quad \text{for } \ell = 1, \dots, L.$$

The empirical or other type of smoothed estimators for the marginal and joint survivor functions  $S_{D, \ell}$ ,  $S_{\bar{D}, \ell}$ ,  $S_{D, \ell_1, \ell_2}(x_1, x_2)$ , and  $S_{\bar{D}, \ell_1, \ell_2}(y_1, y_2)$  can be used to estimate  $v_X$  and  $v_Y$ . In the simulations and the example, we used the empirical estimators. That is, we estimate  $S_{D, \ell}$  and  $S_{\bar{D}, \ell}$  using the expressions in (8). And we estimate  $S_{D, \ell_1, \ell_2}(x_1, x_2)$ , and  $S_{\bar{D}, \ell_1, \ell_2}(y_1, y_2)$  as follows:

$$\begin{aligned} \hat{S}_{D, \ell_1, \ell_2}(x_1, x_2) &= \frac{1}{\sum_{i=1}^M m_{\ell_1 i}^2} \sum_{i=1}^M \sum_{p_1=1}^{m_{\ell_1 i}} \sum_{p_2=1}^{m_{\ell_2 i}} \sum_{k_1=1}^K \sum_{k_2=1}^K I(X_{\ell_1 i p_1 k_1} > x_1, X_{\ell_2 i p_2 k_2} > x_2), \\ \hat{S}_{\bar{D}, \ell_1, \ell_2}(y_1, y_2) &= \frac{1}{\sum_{j=1}^J n_{\ell_1 j}^2} \sum_{j=1}^J \sum_{q_1=1}^{n_{\ell_1 j}} \sum_{q_2=1}^{n_{\ell_2 j}} \sum_{k_1=1}^K \sum_{k_2=1}^K I(Y_{\ell_1 j q_1 k_1} > y_1, Y_{\ell_2 j q_2 k_2} > y_2). \end{aligned}$$

Thus, when  $\Omega$ 's are AUCs,  $v_X$  is given by

$$v_X = \frac{1}{M m_{\ell_1} m_{\ell_2}} \sum_{1 \leq \ell_1, \ell_2 \leq 2R} \sum_{i=1}^M m_{\ell_1 i} m_{\ell_2 i} \frac{\partial h}{\partial \Omega_{\ell_1}} \frac{\partial h}{\partial \Omega_{\ell_2}} \left( E[I(X_{\ell_1 i p_1 k_1} > Y_{\ell_1 j p_1 k_1}) I(X_{\ell_2 i p_1 k_1} > Y_{\ell_2 j p_1 k_1})] - E[I(X_{\ell_1 i p_1 k_1} > Y_{\ell_1 j p_1 k_1})] E[I(X_{\ell_2 i p_1 k_1} > Y_{\ell_2 j p_1 k_1})] \right),$$

and  $v_Y$  is given by

$$v_Y = \frac{1}{M n_{\ell_1} n_{\ell_2}} \sum_{1 \leq \ell_1, \ell_2 \leq 2R} \sum_{j=1}^J n_{\ell_1 j} n_{\ell_2 j} \frac{\partial h}{\partial \Omega_{\ell_1}} \frac{\partial h}{\partial \Omega_{\ell_2}} \left( E[I(X_{\ell_1 i p_1 k_1} > Y_{\ell_1 j p_1 k_1}) I(X_{\ell_2 i p_1 k_1} > Y_{\ell_2 j p_1 k_1})] - E[I(X_{\ell_1 i p_1 k_1} > Y_{\ell_1 j p_1 k_1})] E[I(X_{\ell_2 i p_1 k_1} > Y_{\ell_2 j p_1 k_1})] \right),$$

## 4. Simulation studies

We report simulation studies to evaluate the finite sample property of the proposed statistics. We simulated both multi-reader multi-test ROC data and longitudinal data. In multi-reader multi-test data, we considered the finite sample performance of the variance expression. More importantly, we compared the simulated powers of the equal weight and the optimal weight introduced in Section 2.2. We expect that the optimal weight results in better power than the equal weight.

# Statistics in Medicine

In longitudinal data we considered the general setting where each subject is diagnosed repeatedly at each time point and the number of repeated measures varies from subject to subject.

## 4.1. Multi-reader multi-test data

In the first simulation study we investigated the finite sample accuracy of the variance expression for multireader multitest data. We let  $m_{\ell i} = n_{\ell j} = 1$ ,  $R = 3$ , and  $L = 2$ . We simulated 1000 datasets under multivariate normal and lognormal distributions:

1.  $X \sim N(\mu_X, \Sigma_X)$  and  $Y \sim N(\mu_Y, \Sigma_Y)$ , where  $\mu_X = (1, \dots, 1)$ ,  $\mu_Y = (0, \dots, 0)$  and  $\Sigma_X = \Sigma_Y$  is the variance-covariance matrix with diagonal elements  $(1, 1.5, 2, 1, 1.5, 2)$  and correlation coefficient,  $\rho$ ;
2.  $X \sim \text{LogNormal}(\mu_X, \Sigma_X)$  and  $Y \sim \text{LogNormal}(\mu_Y, \Sigma_Y)$ .

From simulated data we used the proposed statistic in Section 2.2,  $\hat{\Delta}_m = \sum_{r=1}^3 (\hat{\Omega}_r - \hat{\Omega}_{R+r})/R$  to estimate the AUC by defining the weight function  $W(u) = 1$ , for  $0 < u < 1$ , and the pAUC by defining  $W(u) = 1$ , for  $0 < u < 0.6$ ; 0 otherwise. A 95% confidence interval for  $\hat{\Delta}_m$  was obtained using the variance expression derived in (13). Table 1 shows biases, square root of mean squared errors (RMSE), and simulated coverage of confidence intervals. It is clear from the table that coverage levels are close to the nominal level, and biases for comparing AUCs or pAUCs are close to zero. This shows good performance of our estimator and associated asymptotic results.

In the second simulation study we compared the performance of the proposed method with the parametric method by [3] and the semiparametric logistic regression method by [4] with regard to estimating the AUC. We used the same setting as the first simulation study except changing  $\mu_X$  to  $(1, 1, 1, 1.5, 2, 2.5)$ . The biases and RMSEs from the three methods are shown in Table 2. The results indicate that the proposed method and the semiparametric method perform much better than the parametric method when the distribution assumptions are violated. They also indicate that the semiparametric method performs as well as the proposed method. This is not surprising as can be seen from the description of the semiparametric method in Section 2 of [4]. The logistic regression fits the regression parameters based on the following equation:

$$\text{logit}(D = 1) = \beta_0 + \beta_1 Z,$$

where  $D$  is the disease status (with 1 being the diseased, and 0 being the non-diseased),  $\beta_0$  and  $\beta_1$  are regression parameters, and  $Z$  is the test result. After the regression parameter estimators,  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , are obtained, the empirical ROC curve is estimated based on the new score,  $\tilde{Z} = \hat{\beta}_0 + \hat{\beta}_1 Z$ . Since the ROC curve is invariant to monotonic transformation, the empirical ROC curve based on the new score remains the same as the empirical ROC curve from the original test results.

In the third simulation study we compared the simulated powers using the optimal weight versus the equal weight. We again let  $m_{\ell i} = n_{\ell j} = 1$ ,  $R = 3$ , and  $L = 2$ . We simulated 1000 datasets under multivariate normal distributions:  $X \sim N(\mu_X, \Sigma_X)$  and  $Y \sim N(\mu_Y, \Sigma_Y)$ , where  $\mu_X = (2, 1, \dots, 1)$ ,  $\mu_Y = (0, \dots, 0)$  and  $\Sigma_X = \Sigma_Y$  is the variance-covariance matrix with diagonal elements  $(1, 1.5, 2, 2, 3, 2)$  and correlation coefficient,  $\rho$ . We selected  $m = n$  in  $(50, 100)$ , and  $\rho$  in  $(-0.1, 0.2, 0.5)$ . For each simulated data, we estimated the weighted differences in (2.2):

$$h(\Omega) = \left( \sum_{r=1}^3 w_r \right)^{-1} \sum_{r=1}^3 [w_r (\Omega_r - \Omega_{3+r})],$$

with both equal weights ( $w_r = 1/3$ ) and the optimal weights given in (7). The AUC was estimated by defining the weight function  $W(u) = 1$ , for  $0 < u < 1$ , and the pAUC was estimated by defining  $W(u) = 1$ , for  $0 < u < 0.6$ ; 0 otherwise. The simulated power was then calculated as the number of rejections out of 1000 simulated datasets. Table 3 shows the



simulated powers for the comparison of AUCs and pAUCs. It is clear that the optimal weights always result in much larger powers than the equal weights.

## 4.2. Longitudinal biomarker data

In this simulation study we generated multivariate log-normal correlated biomarker data. We generated data by taking exponential of multivariate normal data  $\mathbf{X}_i \sim N(\boldsymbol{\mu}_{X,i}, \boldsymbol{\Sigma}_{X,i})$  and  $\mathbf{Y}_j \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{Y,j})$ , where  $\boldsymbol{\mu}_{X,i} = (2, \dots, 2, 1, \dots, 1)$ , and  $\boldsymbol{\Sigma}_{X,i}$  and  $\boldsymbol{\Sigma}_{Y,j}$  are variance-covariance matrices. We let  $L = 2$ ,  $K = 3$ ,  $M = J = (50, 200)$ . To allow various cluster sizes, we let  $m_{\ell ik} = 2$  for the first half of diseased subjects, and  $m_{\ell ik} = 4$  for the other half. For non-diseased subjects, let  $n_{\ell jk} = 5$  for the first half, and  $n_{\ell jk} = 3$  for the other half. We chose  $\boldsymbol{\Sigma}_{X,i} = (1 - \rho)\mathbf{M}_i + \rho\mathbf{1}_i\mathbf{1}_i'$ , where  $\mathbf{M}_i$  is the  $LKm_{\ell ik} \times LKm_{\ell ik}$  identity matrix and  $\mathbf{1}_i$  is the  $LKm_{\ell ik} \times 1$  matrix with all elements 1. Similar setting was applied to define  $\boldsymbol{\Sigma}_{Y,j}$ . Here  $\rho$  gives within-subject correlation. We let  $\rho = 0.4$  for the diseased and  $\rho = 0.3$  for the non-diseased. We simulated 1000 datasets for each sample size, and obtained the estimate of AUC difference between two biomarkers,  $\hat{\Delta}_l$ , and its variance. Table 4 shows biases, square root of mean squared errors (RMSE), and simulated coverage of confidence intervals. This again shows good performance of our estimator for correlated biomarker data.

## 5. An example in the diagnosis of endometriosis

The proposed nonparametric ROC summary statistics are applied in this section to data from a study on endometriosis diagnosis. Endometriosis is a gynecological medical condition in which endometrial-like cells appear and flourish in areas outside the uterine cavity and is typically seen in women at their reproductive ages. It has been estimated that endometriosis occurs in roughly 5%–10% of women. Despite its relatively high prevalence, substantive and methodological challenges exist, including diagnostic proficiency. The Physician Reliability Study, an add-on to the Endometriosis: Natural History, Diagnosis and Outcome (ENDO) Study [19], addressed this issue by investigating whether sequentially added clinical information of a subject can aid in more accurately diagnosing the disease of endometriosis. Detailed study designs of ENDO and PRS can be found in the aforementioned references. For demonstration purpose in this paper, we used review results of 4 physicians (reviewers) in PRS on 150 participants. All 150 participants had recorded operative digital images of their pelvic organs and descriptive drawings and notes, both from surgeons who conducted the laparoscopies on these women in ENDO study. The reviewers conducted their reviewing and diagnosis under two modalities. Modality one corresponds to the setting where the reviewers are presented with participants' digital video/images while modality two corresponds to the setting where both digital video/images and surgeon's reports (drawings and notes) are presented. For each participant under each modality, the reviewer answered a series questions on what they observe from the clinical information. These answered were later used to derive the rASRM scores [20] which we used as the diagnostic outcomes in this paper. The visualized diagnosis from the original ENDO study of these participants were used as the gold standard.

For the first modality, the estimated AUCs are (0.71, 0.75, 0.63, 0.76) for the four reviewers; the corresponding numbers are (0.83, 0.85, 0.75, 0.87) for the second modality. With equal weights  $w_r = 1/4, r = 1, \dots, 4$ , the  $\Delta$ -statistic is  $\hat{\Delta}_m = -0.1145$ , and its variance estimate is 0.0007475. We used (7) to obtain the optimal weights  $(w_1, w_2, w_3, w_4) = (298.08, 401.16, 176.88, 560.48)$ . Using these weights, the  $\Delta$ -statistic is given by  $\hat{\Delta}_m = -0.1115$ , and its variance estimate is 0.0006961. This indicates that the  $\Delta$ -statistic is more precisely estimated by using the optimal weights. **The two-sided  $p$ -value using the optimal weights is  $2.36 \times 10^{-5}$ , which is slightly smaller than the  $p$ -value  $2.82 \times 10^{-5}$  using equal weights.** The two-sided  $p$ -values based on both sets of weights are both close to zero, which indicates that these physicians are able to give more precise diagnosis on endometriosis by reviewing both digital images and surgeons' descriptive reports.

## 6. Discussion

The proposed methods in the paper are nonparametric and can be applied to evaluate and compare diagnostic markers in the multireader multitest data and the longitudinal data. As illustrated in the simulation studies and the example, the proposed weighted method in the multireader multitest data tends to have a larger power than the existing methods. We also conducted simulation studies to investigate the finite sample performance of the proposed method in the longitudinal data setting. More complex correlated data in which both normal and abnormal locations may occur in the same subject have been considered in [21] and [22]. How to extend the proposed statistics to such a data setting is a future research topic.

**As pointed out by a reviewer, the proposed method is based on the empirical distribution estimators, and may not allow more complicated dependencies of observations in longitudinal data. For example, in the case of autoregressive dependencies, empirical estimators could not converge to target probabilities, especially when autoregression coefficients are greater than one. More research is merited to extend the proposed method in this direction.**

## Acknowledgement

The authors would like to thank an associate editor and two referees for their constructive comments and suggestions. The project described here was supported in part by Award Number R15CA150698 from the National Cancer Institute under the American Recovery and Reinvestment Act of 2009 and by Award Number H98230-11-1-0196 from the National Security Agency. The work was also supported in part with funding from the American Chemistry Council and the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

## References

1. Zhou XH, McClish DK, Obuchowski N. *Statistical Methods in Diagnostic Medicine*. Wiley: New York, 2002.
2. Zou K. Comparison of correlated receiver operating characteristic curves derived from repeated diagnostic test data. *Academic Radiology* 2001; **8**(3):225–233.
3. Molodianovitch K, Faraggi D, Reiser B. Comparing the areas under two correlated ROC curves: parametric and non-parametric approaches. *Biometrical Journal* 2006; **48**:745–757.
4. Copas JB, Corbett P. Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika* 2002; **89**(2):315–331.
5. DeLong ER, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 1988; **44**:837–845.
6. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics* 1997; **53**:567–578.
7. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; **12**:387–415.
8. Zhang D, Zhou X, Freeman D, Freeman J. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. *Statistics in Medicine* 2002; **21**(5):701–715.
9. Baker S, Pinsky P. A proposed design and analysis for comparing digital and analog mammography: special receiver operating characteristic methods for cancer screening. *Journal of The American Statistical Association* 2001; **96**:421–428.
10. Li J, Fine JP. Weighted area under the receiver operating characteristic curve and its application to gene selection. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2010; **59**(4):673–692.
11. Obuchowski N, Rockette H. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. *Communications in Statistics–Theory and Methods* 1995; **24**(2):285–308.

12. Song X, Zhou XH. A marginal model approach for analysis of multi-reader multi-test receiver operating characteristic (ROC) data. *Biostatistics* 2005; **6**(2):303–312.
13. Lee MLT, Rosner BA. The average area under correlated receiver operating characteristic curves: A nonparametric approach based on generalized two-sample wilcoxon statistics. *Applied Statistics* 2001; **50**(3):337–344.
14. Wei LJ, Johnson WE. Combining dependent tests with incomplete repeated measurements. *Biometrika* 1985; **72**(2):359–364.
15. Yang Y, Jin Z. Combining dependent tests to compare the diagnostic accuracies: non-parametric approach. *Statistics in Medicine* 2006; **25**(7):1239–1250.
16. Etzioni R, Pepe M, Longton G, Hu C, Goodman G. Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Medical Decision Making* 1999; **19**:242–251.
17. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989; **76**:585–592.
18. Li G, Zhou K. A unified approach to nonparametric comparison of receiver operating characteristic curves for longitudinal and clustered data. *Journal of the American Statistical Association* 2008; **103**:705–713.
19. Buck Louis GM, Hediger ML, Peterson CM, Croughan M, Sundaram R, Stanford J, Chen Z, Fujimoto VY, Varner MW, Trumble A, *et al.*. Incidence of endometriosis by study population and diagnostic method: the endo study. *Fertility and sterility* 2011; **96**(3):360–365.
20. American Society For Reproductive Medicine. Revised american society for reproductive medicine classification of endometriosis: 1996. *Fertility and Sterility* 1997; **67**:817–821.
21. Werner C, Brunner E. Rank methods for the analysis of clustered data in diagnostic trials. *Computational Statistics & Data Analysis* 2007; **51**(10):5041–5054.
22. Konietzschke F, Brunner E. Nonparametric analysis of clustered data in diagnostic trials: Estimation problems in small sample sizes. *Computational Statistics & Data Analysis* 2009; **53**(3):730–741.
23. Serfling RJ. *Approximation theorems of mathematical statistics*. Wiley: New York, 1980.

## Appendix: Derivation of variance expression of $\Delta_h$

Assume that  $S_{D,\ell}$  and  $S_{\bar{D},\ell}$  have continuous and positive derivatives,  $S'_{D,\ell}$ , and  $S'_{\bar{D},\ell}$ . Suppose that  $M/m_\ell \rightarrow \alpha_\ell$ ,  $M/n_\ell \rightarrow \beta_\ell$ ,  $M/J \rightarrow \lambda$ ,  $\sum_{i=1}^M m_{\ell_1 i} m_{\ell_2 i} / M^2 \rightarrow \eta_{\ell_1, \ell_2}^X$ , and  $\sum_{j=1}^J n_{\ell_1 j} n_{\ell_2 j} / M^2 \rightarrow \eta_{\ell_1, \ell_2}^Y$ , as  $M, J \rightarrow \infty$ . Assume that  $\alpha_\ell$ ,  $\beta_\ell$ ,  $\eta_{\ell_1, \ell_2}^X$  and  $\eta_{\ell_1, \ell_2}^Y$  are finite numbers. In addition, assume that the function  $h$  has continuous partial derivatives of order 2 at each point of an open set  $(\Omega - \epsilon, \Omega + \epsilon)$ , for  $\epsilon > 0$ . Let

$$r_\ell(u) = S'_{D,\ell} \{S_{\bar{D},\ell}^{-1}(u)\} / S'_{\bar{D},\ell} \{S_{D,\ell}^{-1}(u)\}, \quad \text{for } \ell = 1, \dots, L,$$

where  $S'_{D,\ell}$  and  $S'_{\bar{D},\ell}$  are the first derivatives of  $S_{D,\ell}$  and  $S_{\bar{D},\ell}$ , respectively.

The asymptotic normality of  $\hat{\Omega}$  is derived using results from [18], which gives that for markers  $1, \dots, L$ ,

$$\sqrt{M} \begin{pmatrix} \widehat{ROC}_1(u) - ROC_1(u) \\ \widehat{ROC}_2(u) - ROC_2(u) \\ \vdots \\ \widehat{ROC}_L(u) - ROC_L(u) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \sqrt{\alpha_1} \mathbb{U}_{1,1} [S_{D,1} \{S_{\bar{D},1}^{-1}(u)\}] - \sqrt{\beta_1} r_1(u) \mathbb{U}_{2,1}(u) \\ \sqrt{\alpha_2} \mathbb{U}_{1,2} [S_{D,2} \{S_{\bar{D},2}^{-1}(u)\}] - \sqrt{\beta_2} r_2(u) \mathbb{U}_{2,2}(u) \\ \vdots \\ \sqrt{\alpha_L} \mathbb{U}_{1,L} [S_{D,L} \{S_{\bar{D},L}^{-1}(u)\}] - \sqrt{\beta_L} r_L(u) \mathbb{U}_{2,L}(u) \end{pmatrix}$$

where  $\mathbb{U}_{1,\ell}$  and  $\mathbb{U}_{2,\ell}$  are limiting Gaussian processes. Therefore, after some calculation, it follows that

$$\sqrt{M}(\hat{\Omega} - \Omega) \xrightarrow{d} N_L(\mathbf{0}, \Sigma = \Sigma_1 + \Sigma_2), \quad (\text{A.1})$$

where the  $\{\ell_1, \ell_2\}$  element in  $\Sigma_1$  is given by

$$\alpha_{\ell_1} \alpha_{\ell_2} \eta_{\ell_1, \ell_2}^X \int_0^1 \int_0^1 [S_{D,\ell_1, \ell_2} \{S_{\bar{D},\ell_1}^{-1}(s), S_{\bar{D},\ell_2}^{-1}(t)\} - S_{D,\ell_1} \{S_{\bar{D},\ell_1}^{-1}(s)\} S_{D,\ell_2} \{S_{\bar{D},\ell_2}^{-1}(t)\}] dW(s) dW(t), \quad (\text{A.2})$$

and the  $\{\ell_1, \ell_2\}$  element in  $\Sigma_2$  is

$$\lambda\beta_{\ell_1}\beta_{\ell_2}\eta_{\ell_1,\ell_2}^y\int_0^1\int_0^1r_{\ell_1}(s)r_{\ell_2}(t)[S_{\bar{D},\ell_1,\ell_2}\{S_{\bar{D},\ell_1}^{-1}(s),S_{\bar{D},\ell_2}^{-1}(t)\}-st]dW(s)dW(t). \tag{A.3}$$

The Taylor expansion of  $\hat{\Delta}$  at  $\Omega$  gives

$$\hat{\Delta}_h-\Delta_h\overset{d}{\rightarrow}(\hat{\Omega}-\Omega)'\nabla h(\Omega), \tag{A.4}$$

where  $\nabla h(\Omega)$  is the gradient of  $h$  evaluated at  $\Omega$ . Since the asymptotic variance of the right hand side in (A.4) is given by

$$\nabla h(\Omega)'var(\hat{\Omega}-\Omega)\nabla h(\Omega).$$

It follows that

$$var(\hat{\Delta}_h-\Delta_h)\overset{p}{\rightarrow}\sum_{\ell_1,\ell_2}\frac{\partial^2h^2}{\partial\Omega_{\ell_1}\partial\Omega_{\ell_2}}cov(\hat{\Omega}_{\ell_1}-\Omega_{\ell_2},\hat{\Omega}_{\ell_1}-\Omega_{\ell_2}). \tag{A.5}$$

Using the covariance structures in (A.2) and (A.3) in (A.5), we can then obtain the asymptotic normality of  $\hat{\Delta}_h$  by combining (A.1) with the Cramer-Wold device [23].

**Table 1.** Bias, RMSE and coverage for simulated multi-reader multi-test data

	$\rho$	M (J)	AUC			pAUC		
			Bias (in %)	RMSE	Coverage	Bias (in %)	RMSE	Coverage
Norm	-0.1	50	8.01E-02	0.0359	91.94%	3.17E-02	0.0304	92.52%
		100	3.43E-02	0.0483	89.47%	7.99E-02	0.0404	91.99%
		200	-1.93E-01	0.0481	92.18%	-1.00E-01	0.0396	94.40%
	0.2	50	-8.21E-02	0.0258	91.66%	-1.01E-01	0.0217	93.70%
		100	1.31E-01	0.0348	89.87%	1.03E-01	0.0296	91.20%
		200	-1.32E-01	0.0343	92.50%	-1.21E-01	0.0297	92.60%
	0.5	50	-6.38E-02	0.0175	94.12%	-2.01E-02	0.0151	95.70%
		100	-2.78E-02	0.0240	92.10%	-5.44E-02	0.0200	93.00%
		200	6.24E-02	0.0239	94.30%	-7.06E-03	0.0209	94.10%
LN	-0.1	50	-5.01E-02	0.0346	91.99%	1.69E-02	0.0354	92.29%
		100	7.77E-02	0.0478	89.21%	5.27E-02	0.0488	89.38%
		200	-1.38E-01	0.0493	91.98%	-8.07E-04	0.0464	92.59%
	0.2	50	-5.86E-02	0.0261	91.82%	-4.46E-02	0.0250	91.42%
		100	7.04E-02	0.0339	90.16%	7.59E-02	0.0352	89.39%
		200	3.88E-02	0.0340	92.40%	4.38E-02	0.0345	92.70%
	0.5	50	-5.39E-02	0.0169	94.43%	-3.60E-02	0.0172	93.93%
		100	-1.02E-01	0.0241	93.00%	-8.00E-02	0.0234	93.20%
		200	-4.62E-02	0.0239	94.40%	-5.02E-02	0.0243	93.80%

Norm denotes the normal distribution; LN denotes the lognormal distribution.

**Table 2.** Bias and RMSE of the proposed, parametric, and semiparametric methods

	$\rho$	M(J)	Proposed Method		Semiparametric Method		Parametric Method	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
Norm	-0.1	50	-0.0140	0.0329	-0.0123	0.0318	-0.0131	0.0326
		100	-0.0126	0.0251	-0.0144	0.0249	-0.0138	0.0246
		200	-0.0136	0.0202	-0.0132	0.0203	-0.0135	0.0198
	0.2	50	-0.0149	0.0247	-0.0155	0.0440	-0.0117	0.0423
		100	-0.0150	0.0331	-0.0139	0.0327	-0.0125	0.0317
		200	-0.0140	0.0451	-0.0147	0.0262	-0.0136	0.0241
	0.5	50	-0.0133	0.0455	-0.0153	0.0456	-0.0168	0.0446
		100	-0.0132	0.0252	-0.0130	0.0327	-0.0151	0.0330
		200	-0.0132	0.0333	-0.0139	0.0258	-0.0121	0.0239
LN	-0.1	50	-0.0152	-0.0158	-0.0122	0.0360	0.0689	0.0779
		100	-0.0131	-0.0129	-0.0120	0.0265	0.0758	0.0814
		200	-0.0131	-0.0145	-0.0127	0.0203	0.0799	0.0833
	0.2	50	-0.0158	0.0446	-0.0139	0.0499	0.0706	0.0817
		100	-0.0120	0.0232	-0.0141	0.0351	0.0754	0.0810
		200	-0.0136	0.0327	-0.0129	0.0249	0.0807	0.0846
	0.5	50	-0.0158	0.0460	-0.0156	0.0498	0.0705	0.0838
		100	-0.0129	0.0255	-0.0120	0.0344	0.0791	0.0877
		200	-0.0145	0.0343	-0.0134	0.0256	0.0826	0.0884

Norm denotes the normal distribution; LN denotes the lognormal distribution.

**Table 3.** Simulated powers for comparing tests

$\rho$	AUC			
	Equal Weight		Optimal Weight	
	M=J=50	100	50	100
-0.1	0.507	0.741	0.723	0.932
0.2	0.335	0.541	0.659	0.909
0.5	0.327	0.538	0.703	0.936

  

$\rho$	pAUC			
	Equal Weight		Optimal Weight	
	M=J=50	100	50	100
-0.1	0.156	0.290	0.316	0.599
0.2	0.141	0.212	0.280	0.584
0.5	0.133	0.187	0.266	0.643

**Table 4.** Bias, RMSE and coverage for simulated correlated data

	M (J)	AUC			pAUC		
		Bias (in %)	RMSE	Coverage	Bias (in %)	RMSE	Coverage
Norm	50	-0.1182	1.0266	97.40%	0.0627	0.0184	97.40%
	100	0.0302	2.1682	96.60%	0.0931	0.0128	96.60%
	200	0.0038	1.5226	95.80%	0.0116	0.0090	96.00%
LN	50	-0.0768	0.0143	97.10%	0.0097	0.0125	97.10%
	100	-0.1126	0.0218	96.20%	0.0521	0.0093	96.80%
	200	-0.0445	0.0109	94.90%	0.0317	0.0188	95.00%

Norm denotes the normal distribution; LN denotes the lognormal distribution.